

This is the peer reviewed version of the following article:

Application of data fusion techniques to direct geographical traceability indicators / Silvestri, Michele; Bertacchini, Lucia; Durante, Caterina; Marchetti, Andrea; Salvatore, Elisa; Cocchi, Marina. - In: ANALYTICA CHIMICA ACTA. - ISSN 0003-2670. - STAMPA. - 769:(2013), pp. 1-9. [10.1016/j.aca.2013.01.024]

*Terms of use:*

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

07/05/2024 19:58

(Article begins on next page)

1    **APPLICATION OF DATA FUSION TECHNIQUES TO DIRECT GEOGRAPHICAL**  
2    **TRACEABILITY INDICATORS**

3    Michele Silvestri; Lucia Bertacchini; Caterina Durante, Andrea Marchetti, Elisa Salvatore,  
4    Marina Cocchi,

5

6    Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia,  
7    Via Campi, 183 – 41125 Modena (Italy).

8    \* corresponding author: [marina.cocchi@unimore.it](mailto:marina.cocchi@unimore.it) Tel. +390592055029; Fax: +39059373543

9

10   **Abstract**

11    A hierarchical data fusion approach has been developed proposing Multivariate Curve  
12    Resolution (MCR) as a variable reduction tool.

13    The case study presented concerns the characterization of soil samples of the Modena District. It  
14    was performed in order to understand, at a pilot study stage, the geographical variability of the  
15    zone prior to planning a representative soils sampling to derive geographical traceability models  
16    for Lambrusco Wines. Soils samples were collected from four producers of Lambrusco Wines,  
17    insisting in in-plane and hill areas. Depending on the extension of the sampled fields the number  
18    of points collected varies from three to five and, for each point, five depth levels were  
19    considered.

20    The different data blocks consisted of X-ray powder diffraction (XRDP) spectra, metals  
21    concentrations relative to thirty-four elements and the  $^{87}\text{Sr}/^{86}\text{Sr}$  isotopic abundance ratio, a very  
22    promising geographical traceability marker.

23    A multi steps data fusion strategy has been adopted. Firstly, the metals concentrations dataset  
24    was weighted and concatenated with the values of strontium isotopic ratio and compressed. The  
25    resolved components describe common patterns of variation of metals content and strontium  
26    isotopic ratio. The X-ray powder spectra profiles were resolved in three main components that

27 can be referred to calcite, quartz and clays contributions. Then, an high-level data fusion  
28 approach was applied by combining the components arising from the previous data sets.  
29 The results show interesting links among the different components arising from XRDP, the  
30 metals pattern and to which of these  $^{87}\text{Sr}/^{86}\text{Sr}$  Isotopic Ratio variation is closer. The combined  
31 information allowed capturing the variability of the analyzed soil samples.

32

33 **Keywords**

34 Multivariate Curve Resolution, Hierarchical data fusion, XRDP, soils characterization,  
35 geographical traceability markers.

36

37

38

39

## 40 1. Introduction

41 High-throughput methodologies, megavariable database, fast fingerprinting and profiling  
42 techniques are words that nowadays are commonly present in recent literature. The recent  
43 development of analytical techniques able to describe, in a fast way and from different points of  
44 view, several features of the samples under investigations has resulted, in lot of cases, in the  
45 paradox of having a huge amount of data without an effective interpretation [1, 2]. Data fusion  
46 approaches [3-8] are oriented towards the simultaneous use of the information arising from data  
47 of different nature. The joint evaluation of the analytical results allows, at the same time, to  
48 better describe the investigated system and to answer different questions pertaining to which  
49 information it is expected (it is possible) to gain from the different sets/blocks of data, such as  
50 highlighting the consensus, common and distinctive information carried by each block, and the  
51 linkage among them [8].

52 Data fusion techniques can be classified in three main groups: a) *low-level data fusion* consists of  
53 the simple concatenation of the data of different nature b) *mid-level data fusion* is based on  
54 features extraction or variable selection prior to multivariate analysis c) *high-level or*  
55 *hierarchical data fusion* is based on the concatenation of the scores, extracted by means of  
56 multivariate projection techniques such as PCA, PLS, *etc.* [3, 7, 9] or wavelet transform [4], i.e.  
57 models are built separately on the different data blocks and the derived latent variables (or meta  
58 variables in a broad sense) are fused to obtain a final high-multivariate-model.

59 This last approach could be particularly effective in the case of data blocks, which are difficult to  
60 render comparable/commensurable, i.e. a suitable preprocessing procedure may not be available  
61 or completely solve the issue.

62 To the best of our knowledge, multivariate curve resolution (MCR) methodology has not yet  
63 been used as data reduction technique for extraction of data blocks information in high-level data  
64 fusion. The possibility to obtain chemically meaningful components, e.g. that can be

65 characterized in terms of chemical concentration and spectra profiles, allows a better  
66 understanding and highlighting, in the data fusion process, of the correlation between the  
67 resolved profiles of the different analytical techniques.

68 Most often, in data fusion context, MCR has been used combining the information acquired by  
69 the different analytical techniques in the multi-sets structure [10-12]. This is surely a sound  
70 approach, however it may be non optimal when the data sets to be fused all share the samples  
71 mode and each data block is constituted of different kind of variables, e.g. metal contents and  
72 spectral fingerprint for the same set of samples, but there is not a varying condition for each  
73 sample such as time of measurement, pH, or a second spectral dimension. In other word, when  
74 data augmentation limits to variables concatenation and there is not real replicate information for  
75 the same sample to assist the resolution of the underlying components.

76 Here, we present a case study where MCR was used as variable reduction tools for the  
77 development of hierarchical data fusion model in a study aimed at achieving information about  
78 the geochemical variability of soils samples.

79 In particular, this work is a part of a pilot study belonging to a project concerning assessment of  
80 geographical traceability models for Lambrusco wines of protected denomination of origin  
81 (PDO), a typical food product of the Province of Modena (Italy).

82 Food geographical traceability studies are targeted to establish the correlation between the soils  
83 of origin and the final products, hence, one of the main aspect to face is the representativeness of  
84 the sampling of the territories under investigation. To characterize soils samples the  
85 determination of several metals content together with the isotopic abundance ratio  $^{87}\text{Sr}/^{86}\text{Sr}$  (a  
86 very promising geographical traceability marker) were used jointly with the X-Ray Powder  
87 diffraction profile to obtain information about the inter and intra-site variability, including depth,  
88 of soil samples at few selected locations, as well as to evaluate the link among main  
89 mineralogical phases, metals and isotopic abundance ratio  $^{87}\text{Sr}/^{86}\text{Sr}$ .

90 The MCR data fusion approach was preferred instead of the multiset based one, for the reason  
91 explained above taking into account the great difference among data blocks in terms of number  
92 of variables and measurement scales.

93 Several examples are present in literature for the identification of patterns of variation of  
94 pollutants or metal sources based on MCR [13, 14] whereas it is the first time that an approach  
95 based on multivariate curve resolution is proposed to attempt a partial resolution of XRDP  
96 components. In particular, we were interested on one hand to fully exploit soil samples  
97 differences and similarity, by building a comprehensive model merging the resolved MCR  
98 components from metals and isotopic ratio dataset with the XRDP one, and on the other hand to  
99 focus on the linking relations between the fused data blocks.

## 100 **2. Materials and Methods**

### 101 *2.1. Experimental*

102 The sampling procedure and the analytical determinations concerning the data reported in each  
103 data block have been described in our previous works [15 and references therein]. Here we will  
104 briefly report only the salient information.

#### 105 *2.1.1. Soils Sampling*

106 Production of PDO Lambrusco wines is subjected to stringent regulations [16, 17] allowing  
107 grapes cultivation in the whole district of Modena. The territory of the Province of Modena  
108 varies from in-plain area (centre-north) to moderate hill area in the south.

109 The extension of the area (more than 90 km<sup>2</sup>) and the amount of Lambrusco producers insisting  
110 on it (more than four thousand) made it mandatory to develop a pilot sampling to evaluate, on a  
111 reduced scale, variability of soils in the district, sampling conditions and operating procedures  
112 [15]. Thus, four long chain producers were considered, three of these producers, named here on  
113 as A, B and D are located in in-plain region, where the majority of the production of Lambrusco

114 wines insists, the fourth one, C producer, insists in the hill area. For each producer, depending on  
115 the dimension of the field, from three to five coring where collected. In order to obtain  
116 information about both horizontal and vertical variability, each core was split in five aliquots of  
117 10 cm of length, starting from 10 cm of depth to 60 cm. All depths were analyzed for the hill  
118 field and only lower and upper aliquots for the plain ones for a total of 47 samples.

### 119 *2.1.2 Analytical Determination*

120 X-ray diffraction of powder, XRDP, was carried out by a  $\theta/\theta$  PANalytical X'Pert Powder  
121 diffractometer equipped with a Real Time Multiple Strip (RTMS) detector (PANalytical  
122 X'Celerator). Metals quantification needed a preliminary phase of sample pretreatment. Analytes  
123 were transferred from soils into solution, in order to perform the analytical measurement, by  
124 means of an acid leaching, with concentrated Suprapur® HNO<sub>3</sub>, assisted by microwave. The  
125 quantification of the Ca, Mg, K, Na content was evaluated by means of a F-AAS (SpectrAA  
126 220FS, supplied by Varian, equipped with a sample introduction/dilution system, SIPS 10) . An  
127 inductively coupled plasma mass spectrometer, ICP/qMS, XSeriesII from Thermo Fisher  
128 Scientific (Bremen, Germany), was used for the determination of the following isotopes: <sup>7</sup>Li,  
129 <sup>51</sup>V, <sup>52</sup>Cr, <sup>60</sup>Ni, <sup>63</sup>Cu, <sup>66</sup>Zn, <sup>68</sup>Zn, <sup>71</sup>Zn, <sup>85</sup>Rb, <sup>88</sup>Sr, <sup>109</sup>Ag, <sup>114</sup>Cd, <sup>133</sup>Cs, <sup>137</sup>Ba, <sup>139</sup>La, <sup>140</sup>Ce, <sup>146</sup>Pr,  
130 <sup>149</sup>Sm, <sup>151</sup>Eu, <sup>158</sup>Gd, <sup>163</sup>Dy, <sup>165</sup>Ho, <sup>167</sup>Er, <sup>169</sup>Tm, <sup>172</sup>Yb, <sup>175</sup>Lu, <sup>205</sup>Tl, <sup>208</sup>Pb, <sup>232</sup>Th, <sup>238</sup>U. The  
131 evaluation of isotopic abundance ratio <sup>87</sup>Sr/<sup>86</sup>Sr was achieved by means of a multi-collector high-  
132 resolution ICP/MS Neptune® provided by Thermo Scientific after the separation of the isobaric  
133 interference of rubidium via SPE with Eichrom's Sr Resin (4,4'(5')-di-t-butylcyclohexano 18-  
134 crown-6 crown ether).

### 135 *2.2 Data Analysis*

#### 136 *2.2.1 Multivariate Curve Resolution*

137 MCR is based on bilinear decomposition of the data matrix [18, 19] according to the model:

138  $\mathbf{D} = \mathbf{C} \mathbf{S}^T$

139 The model is calculated by alternating least squares algorithm (MCR-ALS).

140 Since MCR is not an orthogonal decomposition such as PCA, it needs constraints to resolve the  
141 system in a way that the S (spectra) matrix corresponds to a real chemical behavior.

142 Constraints can be applied both to the spectra (S) and the concentration (C) matrices, in order to  
143 reduce the rotational ambiguity of the model since MCR-ALS has not a unique solution.

144 Constraints can be considered as the translation in mathematical formulae of a characteristic of  
145 the investigated system.

146 Two types of constraints can be implemented in an MCR model: i) soft constraints such as non-  
147 negativity, unimodality, selectivity and closure constraints that allow reducing rotational  
148 ambiguities; ii) hard constraints, based on physicochemical models able to describe the system  
149 under investigation such as kinetic or equilibrium model are able to reduce in the same time both  
150 rotational and intensity ambiguities.

151 In the resolution of a chemical system, non-negativity constraints are very common. Furthermore  
152 within the family of constraints [20] other usefully adopted for the reduction of the ambiguities  
153 are: unimodality (i.e. the resolved profiles are imposed to have only a maximum), closure (i.e.  
154 the total amount of the species within the system is constant) and selectivity (i.e. imposition of  
155 the presence or absence of a species in a mixture or a region of the spectrum).

156 Here, we adopted soft constrains, such as non-negativity constraints both for concentrations and  
157 spectral profiles. Constrains motivation and application details for each data block are reported  
158 in the section 3, Discussion.

159

160 *2.2.2 Datasets and Preprocessing*



161 As pointed out in the Introduction, we decided to follow a hierarchical data fusion approach,  
162 analyzing separately by multivariate curve resolution methodology the two data sets and then  
163 merging the extracted concentration profiles by each of them. In fact, the way in which variables  
164 are weighted is one of the most important tasks in data-fusion. In particular, in the case of fusion  
165 of blocks with very different number of variables and measurement scales, as in this case, low-  
166 level data fusion could be difficult. Forcing a block of few variables to have the same variance as  
167 a more numerous one can lead to a concatenated dataset having the majority of variable with  
168 “intensities” very low and few ones with high values. This is more troublesome in the case of the  
169 use of multivariate curve resolution instead of orthogonal methods, such as PCA.

170 The data fusion approach we adopted addresses two purposes: meaningful components are  
171 extracted and the two data blocks are made comparable as more or less the same number of  
172 variables (resolved component) each one having a high variance with respect to the original data  
173 are concatenated.

174 The whole data fusion process is illustrated in figure 1 and described here after.

175 The collected diffractograms consist of 6882  $2\theta$  intensities (from  $0^\circ$  to  $120^\circ$  on  $2\theta$  scale).  
176 The last part of the signals was cut at  $79.99\ 2\theta$  and diffractograms were then preprocessed in  
177 order to reduce noise and background effects and to minimize horizontal shift [15]. The signals  
178 were finally arranged in a  $47 \times 4488$  matrix called “*XRDP dataset*”. By means of MCR-ALS  
179 three concentration profiles were extracted and arranged in a  $47 \times 3$  matrix call “*MCR-XRDP*  
180 *dataset*”.

181 On the other side, the concentrations ( $\text{mg kg}^{-1}$ ) of the 34 metals were scaled to unit variance  
182 (without centering) and merged together with the isotopic abundance ratio values. The obtained  
183 matrix, namely “*Met-I.R. dataset*” of dimension  $47 \times 35$  was block-scaled in a way that the 15%  
184 of the total variance correspond to  $^{87}\text{Sr}/^{86}\text{Sr}$  in order to assign to this important primary indicator  
185 a value six times higher with respect to each other metals concentration. Strontium abundance

186 isotopic ratio is an important parameter in geochemistry and geochronology, its value depends  
187 on several factors such as the age of formation of soils and the initial concentration of  $^{86}\text{Sr}$ ,  $^{87}\text{Sr}$   
188 and  $^{87}\text{Rb}$ .  $^{87}\text{Rb}$  is not a stable isotope and decays to  $^{87}\text{Sr}$  with an half-life time ( $t_{1/2}$ ) of more than  
189 ten to the tenth years. Given all these properties, we decided to enhance the importance of this  
190 variable with respect to all the others.

191 Four concentration profiles were extracted with MCR-ALS and arranged in a 47x4 dataset called  
192 "*MCR- Met-I.R. dataset*"

193 Finally the two concentration profiles matrices were concatenated and block-scaled in order to  
194 give the same variance to each block and then MCR-ALS were applied giving the bilinear  
195 decomposition of the arranged dataset "*DF MET-I.R.-XRDP Dataset*" based on three  
196 components.

### 197 *2.2.3 Software*

198 Multivariate Curve Resolution was carried out by MCR-ALS GUI  
199 ([http://www.ub.edu/mcr/web\\_mcr/mcrals.html](http://www.ub.edu/mcr/web_mcr/mcrals.html)). Arrangement of dataset and fusion of data was  
200 obtained by homemade routine written with MATLAB (Mathworks MA, USA) and PLS  
201 Toolbox 6.0 (distributed by Eigenvector Research Inc WA, USA).

202

## 203 **3 Results and Discussion**

204 This section is divided in three parts describing the application of MCR-ALS to the different  
205 datasets: 3.1 results on "*XRDP dataset*", 3.2 results on "*MET-I.R. dataset*", 3.3 results on "*DF*  
206 *MET-I.R.-XRDP Dataset*"

### 207 *3.1 XRDP Dataset*

208 Diffractograms of soils powders are very complex signals characterized by a high number of  
209 peaks of different intensities. The number of mineralogical phases contributing to the whole  
210 signal (chemical rank) depends on the complexity of the soils. In this work the application of  
211 MCR-ALS on the XRDP signals is not oriented to a complete quantitative resolution of the  
212 system under investigation but it is proposed as tool for the extraction of the meaningful  
213 information able to characterize the investigated soils. In fact, here the aim is not the  
214 quantification of the single mineralogical phases, but rather a qualitative identification of the  
215 phases responsible of the main sources of variability for soil samples differentiation. The use of a  
216 rank deficient model for the description of a system leads to a not complete resolution of the  
217 components that results in overlapped spectra profiles. If few major components are presents and  
218 able to characterize adequately the system under investigation, as in this case, the choice of a  
219 rank deficient model is sufficient but the so called “pure component” probably contains part of  
220 the information of other components not considered in the curve resolution process.

221 Several attempts were tried for the application of MCR-ALS on the “*XRDP dataset*”, varying  
222 the number of components, initial estimates, constraints and normalization strategies.

223 Preliminarily a MCR model with three components was chosen, based on singular value  
224 decomposition (SVD) results. Initial estimates were determined by SIMPLISMA [21] on spectra,  
225 non-negativity constraints were applied both for concentration and spectra, and spectra profiles  
226 were height normalized. Two of the three resolved spectra profiles (not reported) by this  
227 preliminary model resulted to be very similar to signals of the pure compounds calcite (the most  
228 stable polymorph of calcium carbonate) and quartz (silicon dioxide). In the third spectra profile  
229 most of the bands were present in the low  $2\theta$  region and many clays related peaks could be  
230 identified in the middle region of the spectrum.

231 Quartz, clays and calcite, in order of abundance, are the main constituents covering the majority  
232 of the compositional profile of soils insisting in the Modena district, so it seems reasonable to

233 refine the model described above implementing selectivity constraints in order to reduce the  
234 rotational ambiguity of the system.

235 In particular, few selected peaks were forced to be modeled only by one of the component on the  
236 basis of the XRPD signal of the pure quartz and calcite respectively, as retrieved by the database  
237 powder diffraction files version 4 (ICDD). The XRPD region corresponding to  $50.1\ 2\theta$  was  
238 imposed to be selective for quartz. The regions corresponding to  $29.5\ 2\theta$ ,  $39.4\ 2\theta$ ,  $48.5\ 2\theta$  were  
239 imposed to be selective for the “calcite” component, in these regions calcite responds, and not  
240 any other interfering species, or at least a contribution may be given only by species that are  
241 extremely rare in the investigated territory.

242 The three components model built implementing selectivity constraints explains 92% of the total  
243 variance with a lack of fit with respect to PCA of 1.1%. In figure 2 are reported the three  
244 resolved spectra profiles, The tentative identification of the three resolved profiles, as clays,  
245 calcite and quartz does not pretend we are referring to completely resolved profiles for these pure  
246 components, as pointed out above, we are aware of having chosen to fit a rank deficient model  
247 leading to not complete resolution. However, the three MCR components, as shown by the  
248 results obtained by the MCR model not forced with selectivity constraints and then confirmed by  
249 the constrained MCR model, are quite close to the pure spectra of these mineralogical phases,  
250 that represent the main constituents of soils of the Province of Modena. Hence, for convenience  
251 the terms “clays”, “calcite” and “quartz” will be associated to the resolved spectra profiles.

252 In figure 3 is highlighted a region of the spectra in which is clear the good resolution of other  
253 peaks belonging to the calcite component that were not forced by the applied selectivity  
254 constraints.

255 In order to inspect the information about samples, pertaining to the concentration matrix C, a  
256 scatter plot representation of its columns is proposed in figure 4. As MCR is not an orthogonal  
257 decomposition, this representation, while offering an easier way, with respect to separate bar

258 graphs for each C column, to formulate considerations about the similarities or differences of  
259 samples on the basis of two components at time, does not have the same meaning as samples  
260 distance in PCA spaces. However, the coordinates of samples in the concentration space of the  
261 MCR factors are directly related to the concentration and should have a more directly  
262 interpretable meaning than the distances between samples in the space of orthogonal factors.

263 The most striking observation that emerges in figure 4 is the separation of the hill samples  
264 (producer C) that are separated for low concentration values for the clays component with  
265 respect to all the in-plane samples (producers A, B and D). Moreover, the hill samples are split  
266 in two groups due to a different amount of both calcite and quartz. The distinction in two clusters  
267 of the hill samples is in agreement with a certain degree of soils variability at the site of producer  
268 C (hill area) that was also noticed during the on field sampling (15) and confirmed by  
269 preliminary results from texture analysis of the same soil samples where the whole content of  
270 sand, clay, silt and  $\text{CaCO}_3$  were determined. The distribution of these components in the  
271 different holes shows a great complexity and heterogeneity of these soils. Thus the intra site  
272 location of the collected samples may reflect the presence of a more calcareous soil fraction with  
273 respect to others characterized by a high amount of quartz. At variance, all in-plain samples are  
274 slightly more homogeneous and only the samples of the producer A present a lower amount for  
275 the calcite component and higher for the quartz ones with respect to the samples of the producers  
276 B and D.

### 277 3.2 MET-I.R. Dataset

278 The dataset “MET-I.R. dataset” containing the concentrations of the 34 metals determined and  
279 the isotopic abundance ratio  $^{87}\text{Sr}/^{86}\text{Sr}$  was pretreated as described in the previous section, also in  
280 this case several attempts were tried for the selection of the best parameters for the application of  
281 MCR-ALS.

282 A four components model was calculated using SIMPLISMA on concentration direction for the  
283 evaluation of initial estimates, applying non-negativity in both concentration and “spectra”  
284 (variables) direction (99% of total variance explained, lack of fit with respect to PCA 0.8%).

285 The resolved component profiles (variables mode) are reported in figure 5. Each profile  
286 represents a common pattern of variation of the investigated soils resolved by MCR-ALS. The  
287 information that can be extracted from these profiles is a characterization of the samples on the  
288 basis of groups of metals that vary likewise for all the soils analyzed. It is possible to denote that  
289 some variables present high values for almost all the components (Cr, U, Ni).

290 On the other hand, other variables are not present in some components, for example Calcium  
291 presents values near or equal to zero for the first, second and fourth component and the highest  
292 one for the third one, as magnesium, which shows the same behavior. Sodium and potassium  
293 have high contribution in the second and third components but not in the other two, the same  
294 trend is observed for rare earths elements in the first, second and fourth component. Isotopic  
295 abundance ratio of  $^{87}\text{Sr}/^{86}\text{Sr}$  presents high values for the first component and low or very low for  
296 all the other ones.

297 Considering the samples concentration profiles obtained for the MET-I.R. dataset, which are  
298 reported in figure 6, they share similarities with respect to the trends shown above in figure 4. In  
299 particular the second component arising from the MET-I.R. model is associated to a distribution  
300 of the samples, as regards concentrations values, very similar to the clays component of the  
301 XRDP dataset. The same consideration can be extended for the third MET-I.R. component and  
302 the calcite ones. The first component, on the other hand, presents higher values for all the hill  
303 samples with respect to the in-plane ones. The fourth component is not able to highlight  
304 differences on the groups of soils, only one hill sample is present at very high values indicating a  
305 great amount for this soil of one or more of the variables with high contribution on the fourth  
306 component (e.g. cadmium).

### 307 3.3 DF MET-I.R.-XRDP Dataset

308 The three resolved concentration profiles from XRDP model were concatenated with the four  
309 from the analysis of metal and isotopic ratio dataset in a way that the variance explained by each  
310 block was the same, since the aim the data fusion process is to extract the correlation between  
311 the two blocks of variables in order to better explain the variability of the soils samples.

312 The resulting “DF MET-I.R.-XRDP dataset” of dimensionality 47x7 was evaluated by a three  
313 component MCR-ALS model, using SIMPLISMA on concentration direction for the evaluation  
314 of initial estimates and applying non-negativity constraints for both concentration and “spectra”  
315 (variables mode). The model is characterized by 95.9% of variance explained, lack of fit with  
316 respect to PCA 0.3%.

317 In figure 7a, 7b and 7c are shown the bar graphs of the three MCR components respectively, for  
318 the S matrix (variables mode). To better describe the linkage among the components extracted by  
319 the different data blocks in figure 7d the three-dimensional scatter plot is shown. In particular, it  
320 can be observed the grouping of the quartz component from the XRDP model and the first  
321 component from MET-I-R that get the highest value on the first resolved component of the fused  
322 dataset. The first MET-I.R. component is principally related to the values of the isotopic  
323 abundance ratio of  $^{87}\text{Sr}/^{86}\text{Sr}$ , *U*, *Cr*, (that presents the maximum value, figure 5 top left) and to  
324 the contribution of all the rare earths patterns. Thus, in the investigated soil samples to higher  
325 quartz content corresponds as well a higher isotopic abundance ratio  $^{87}\text{Sr}/^{86}\text{Sr}$ .

326 On the second resolved component of the fused dataset, clays (XRDP data set) and the second  
327 and fourth of the MET-I.R. dataset show the highest values. The correlation between one XRDP  
328 component and two components from the metal dataset (mostly influenced by transition metals:  
329 zinc, nickel, cobalt, vanadium, cadmium, monovalent elements such as rubidium, potassium,  
330 sodium, and by the rare earths pattern) highlight the complexity of the possible relation between  
331 the different variables (metals composition and lattice structures) and it can also be due to the not

332 complete resolution of the clays component, that is a combination of several crystalline structure  
333 of different clays (muscovite, illite, chlorite, serpentine) typically presenting the inclusion of  
334 different metals in the lattice.

335 The third component shows the correlation between the calcite component and the third MET-  
336 I.R. component that, as expected, is characterized by the highest contribution of calcium with  
337 respect to all the other spectra profiles and other bivalent elements that commonly are presents in  
338 carbonate soils, such as magnesium, strontium, zinc, nickel and by alkaline metals such as  
339 sodium, potassium, rubidium.

340 In figure 8 are reported the concentration profiles. The distribution of the samples presents  
341 similarities with the separate models presented above. All in-plain samples have the highest  
342 values for the second component (as they showed on the second component in MET-I.R model  
343 and the clays component in the XRDP model) but are more differentiated in the first and third  
344 component. Samples from the A producer present more positive values for the first component  
345 and lower values for the third one with respect to all other in-plain samples. This indicates that  
346 the field of producer A has higher amount of quartz and strontium isotopic ratio, the most  
347 important variable for the second component of the MET-I.R model.

348 Regarding the samples from the hill field, it is clear the greater variability compared to the in-  
349 plain ones. Analyzing the first and third component plot in figure 8b it is evident the separation  
350 in two groups of the hill samples. These components are related to variables from XRDP and  
351 MET-I.R. dataset that appear to be mutually complementary. The presence of a high amount of  
352 one of them is linked to the presence of a lower one for the other component. These results are in  
353 agreement with preliminary information regarding the textures analysis of the same soils, such as  
354 the percentage of sand and silt. This variability is so pronounced that some samples of the same  
355 hole but at different depth present high values for the third component in some cases and lower  
356 for some other.



357 **4. Conclusions**

358 The high-level data fusion approach adopted for the analysis of data of different nature proved to  
359 be a powerful tool for a preliminary inspection of blocks of data characterized by a very different  
360 dimensionality. In this contest, MCR-ALS proved to be a relevant variable reduction mean  
361 because the extracted components bear chemically meaningful information, hence an easier  
362 interpretation of the data in each stage of the data fusion process.

363 When preliminary knowledge about variability and features of samples are not present or not  
364 detailed, as in this case, the possibility of the interpretation of the results on the basis of a  
365 multifold instrumental approach helps the determination of both the characteristics of samples  
366 and of the single results of each block of variables. In particular, the linkage among the metal  
367 profile and some mineralogical phases with the isotopic ratio may furnish a preliminary estimate  
368 of what variability has to be expected depending on field location and thus aids the planning of  
369 soil sampling.

370 The applications of MCR-ALS on very complex data, such as diffractograms, demonstrate the  
371 capability of this tool of achieving good results when used on fingerprinting techniques data in a  
372 way similar to an exploratory analysis.

373 **5. Acknowledgments**

374 This work was supported by the AGER, Agroalimentare e Ricerca, cooperative project between  
375 grant-making foundations under the section “wine growing and producing” project. New  
376 analytical methodologies for varietal and geographical traceability of enological products;  
377 contract n. 2011-0285. We are also grateful to Consorzio Marchio Storico Lambruschi Modenesi  
378 for use of their facilities during sampling procedures.

379 Moreover, we would like to thank Prof. A. De Juan for valuable suggestions.

380

381 **References**

- 382 [1] L. Eriksson, H. Antti, J. Gottfries, E. Holmes, E. Johansson, F. Lindgren, I. Long, T.  
383 Lundstedt, J. Trygg, S. Wold, *Anal Bioanal Chem* 380 (2004) 419-429.
- 384 [2] P. J. Gemperline, *J. Chemometrics* 2007; 21: 507–508
- 385 [3] J. Forshed, H. Idborg, S.P. Jacobsson, *Chemometr Intell Lab*, 85 (2007) 102–109
- 386 [4] Y. Lin, S.D. Brown, *Anal Bioanal Chem*, 380 (2004) 445-452
- 387 [5] A. K. Smilde, M. J. van der Werf, S. Bijlsma, B. J. C. van der Werff-van der Vat, R. H.  
388 Jellema, *Anal Chem*, 77 (2005) 6729-6736
- 389 [6] Y. Nia , Y. Lai, S. Brandes, S. Kokot, *Anal Chim Acta*, 647 (2009) 149–158
- 390 [7] S. Wold, N. Kettaneh, K. Tjessem, *J. Chemometr*, 5 (1996) 463-482
- 391 [8] I. Van Mechelen, A.K. Smilde, *Chemolab* 104 (2010) 83-94
- 392 [9] J. A. Westerhuis, T. Kourti, J. F. MacGregor, *J. Chemometrics*, 12 (1998) 301–321
- 393 [10] E. Peré-Trepat, R. Tauler, *J. Chromatogr A*, 1131 (1-2) (2006) 85-96
- 394 [11] S. Navea, R. Tauler, A. de Juan, *Anal Chem*, 78 (2006) 4768-4778
- 395 [12] S. Mas, R. Tauler, A. de Juan, *J. Chromatogr A*, 1218-51 (2011) 9260-9268
- 396 [13] A. Gredilla, J. M. Amigo, S. Fdez-Ortiz de Vallejuelo, A. de Diego, R. Bro, J. M.  
397 Madariaga, *Anal Methods*, 4 (2012) 676]
- 398 [14] E. Pere-Trepat, A. Ginabreda, R. Tauler, *Chemometr Intell Lab*, 88 (2007) 69-83
- 399 [15] L. Bertacchini, C. Durante, A. Marchetti, S. Sighinolfi, M. Silvestri, M. Cocchi, *Talanta*  
400 (2012) <http://dx.doi.org/10.1016/j.talanta.2012.06.067>
- 401 [16] Decree of 27 July 2009 published on O.J. no. 187 of 13 August 2009.
- 402 [17] CE Regulation no. 813 of 17 April 2000.
- 403 [18] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, *Chemometr Intell Lab*, 76 (2005) 101-110
- 404 [19] S.C. Rutan, A. de Juan, R. Tauler, *Comprehensive Chemometrics – Chemical and*  
405 *Biochemical Data Analysis*, Volume 2, 249-259

- 406 [20] A. de Juan, Y. Vander Heyden, R. Tauler, D.L. Massart, *Anal Chim Acta* 346 (1997) 307-  
407 318
- 408 [21] W. Windig, *Chemom. Intell. Lab. Syst.* (1997), 36, 3–16.  
409

410 **Captions of figures**

411 **Figure 1:** Schematization of the data fusion process

412 **Figure 2:** MCR-ALS resolved spectra profiles of XRDP dataset. a) First component, “clays”. b)  
413 Second Component, “calcite”. c) Third Component, “quartz”.

414 **Figure 3:** Zoom of the 40-49 2theta region showing superimposed the MCR-ALS resolved  
415 profiles for the three XRDP components.

416 **Figure4:** MCR-ALS resolved concentration profiles of XRDP dataset. a) “clays” versus “calcite”  
417 resolved concentration profiles. b) “calcite” resolved versus “quartz” resolved concentration  
418 profiles.

419 **Figure5:** MCR-ALS resolved “spectra” profiles (variables mode) of MET-I.R. dataset.

420 **Figure6:** MCR-ALS resolved concentration profiles of MET-I.R. dataset. a) third component  
421 versus first component; b) second component versus third component; c) first component  
422 versus fourth component.

423 **Figure7:** MCR-ALS resolved “spectra” (variables mode) of the fused dataset DF MET-I.R.-XRDP.  
424 a) First component. b) Second component. c) Third component. d) Scatter plot of three of the  
425 resolved components.

426 **Figure8:** MCR-ALS resolved concentration profiles of the fused dataset DF MET-I.R.-XRDP. a)  
427 Second component versus first component; b) first component resolved versus third  
428 component; c) second component versus third component.

429

Figure 1  
[Click here to download high resolution image](#)

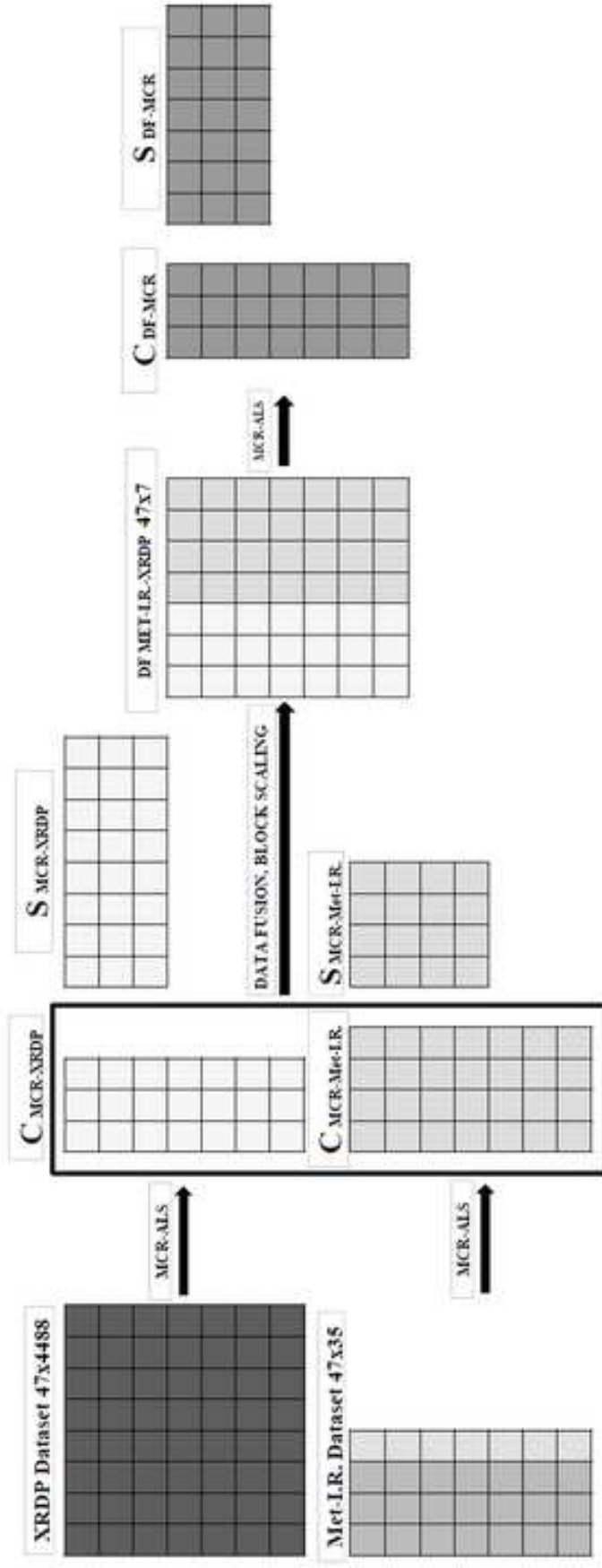


Figure 2  
[Click here to download high resolution image](#)

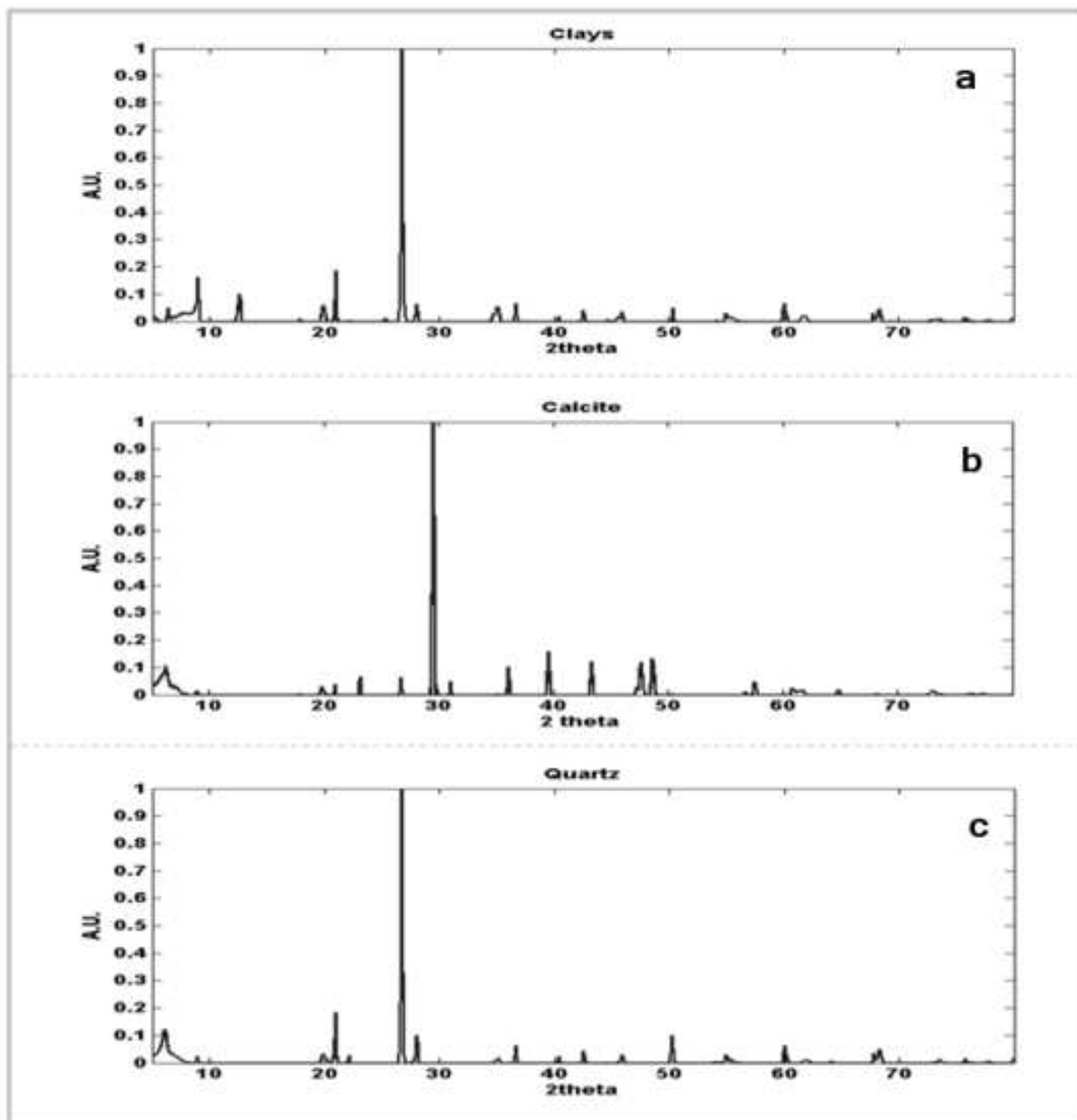


Figure 3  
[Click here to download high resolution image](#)

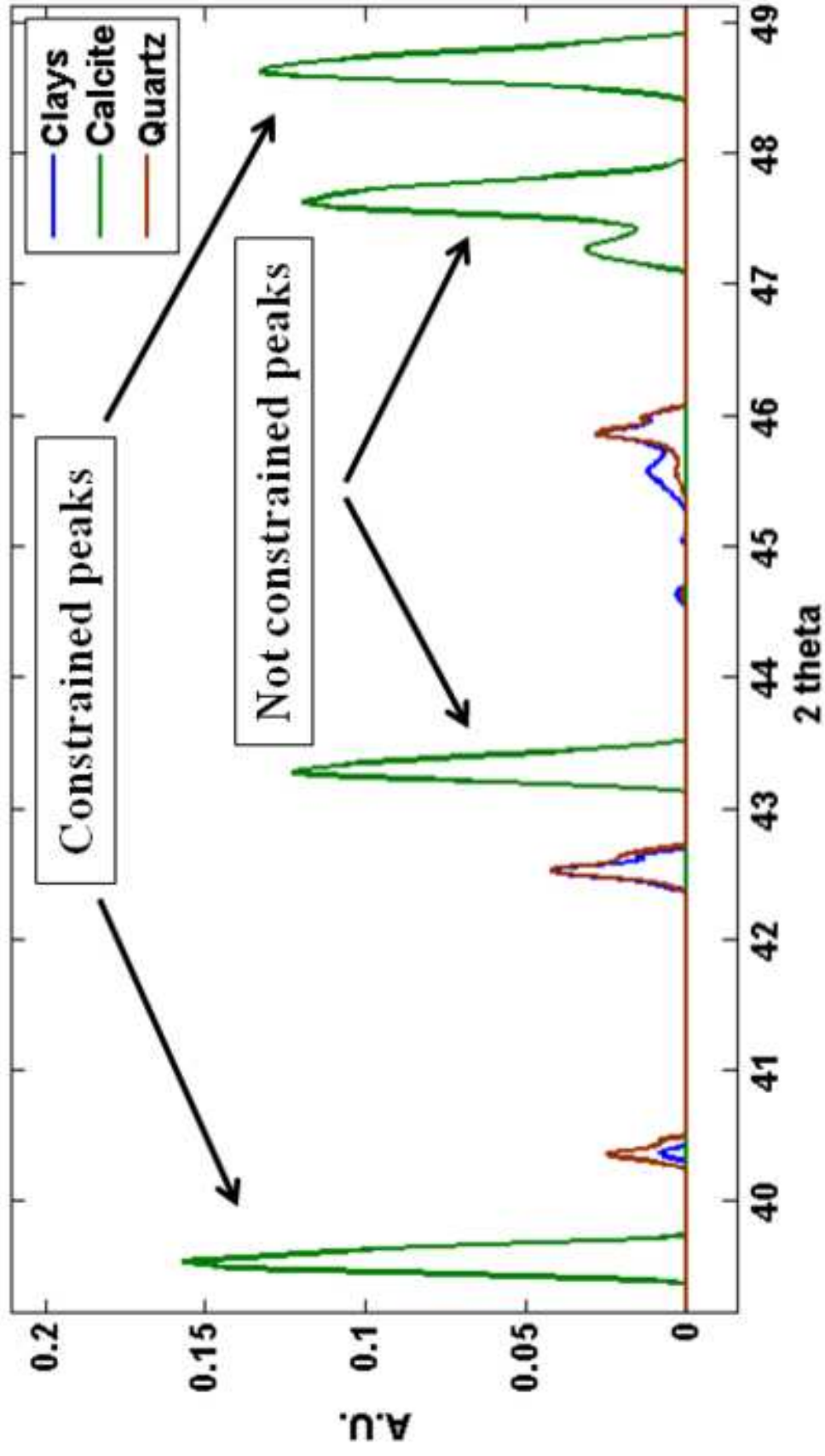


Figure 4a  
[Click here to download high resolution image](#)

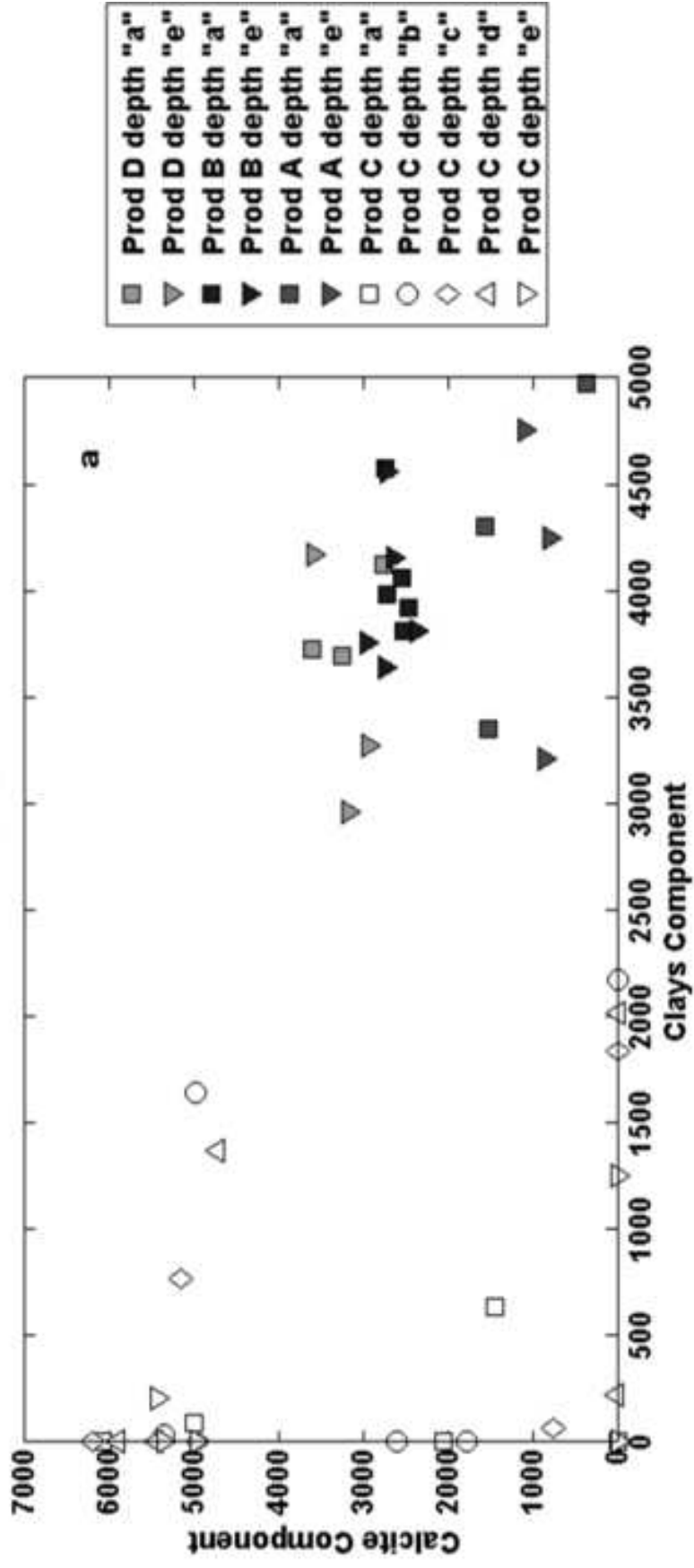




Figure 4b  
[Click here to download high resolution image](#)

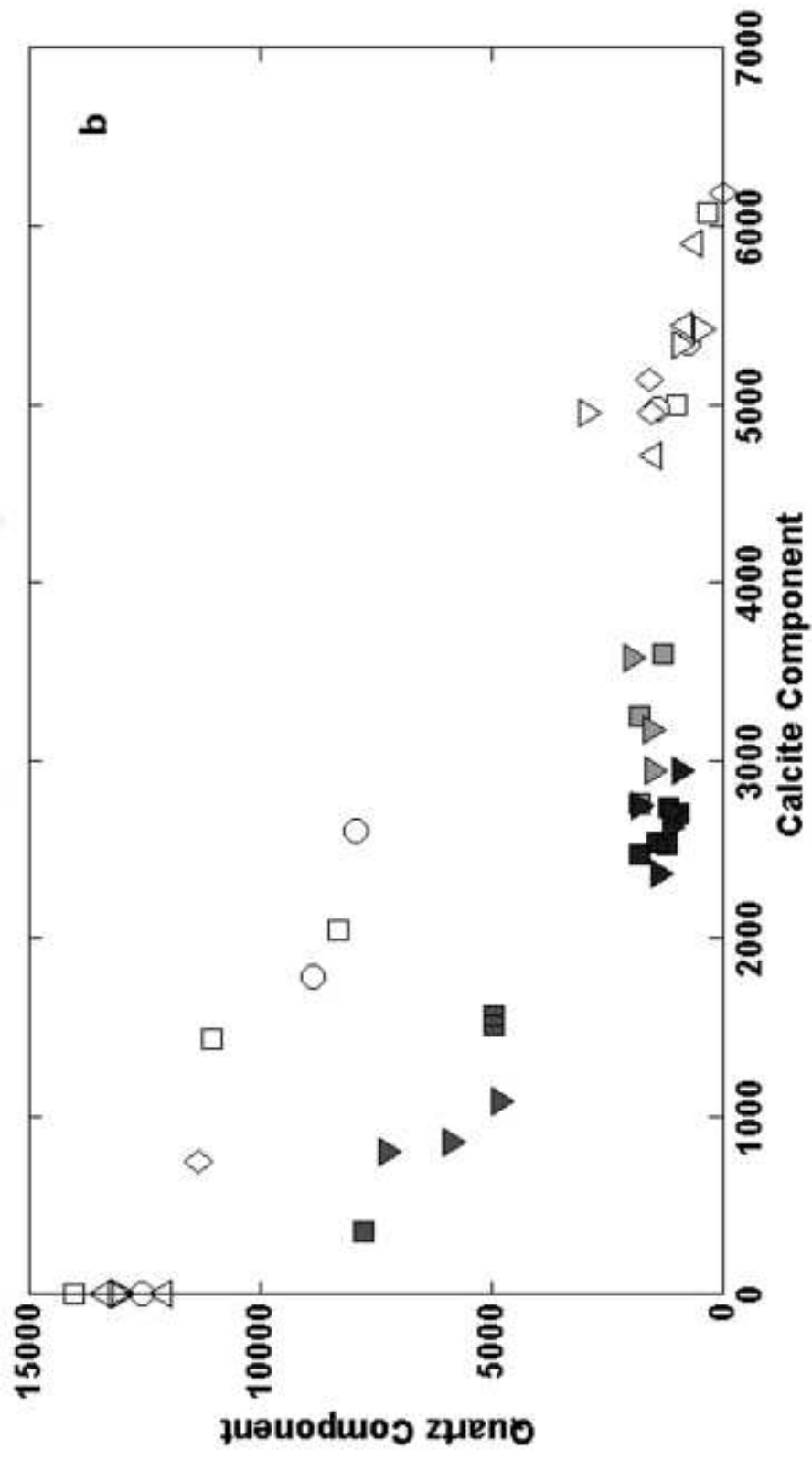


Figure 5  
[Click here to download high resolution image](#)

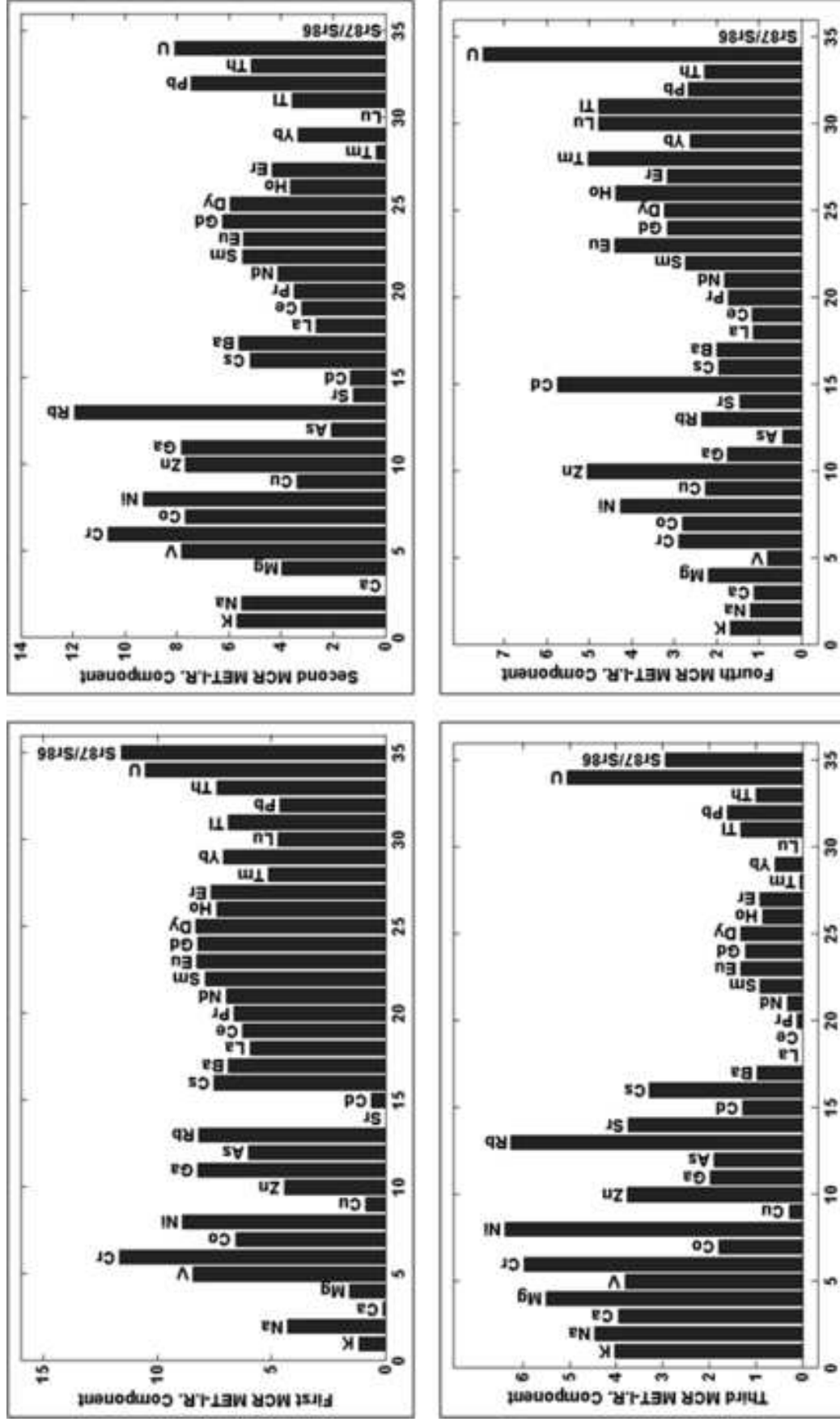


Figure 6a  
[Click here to download high resolution image](#)

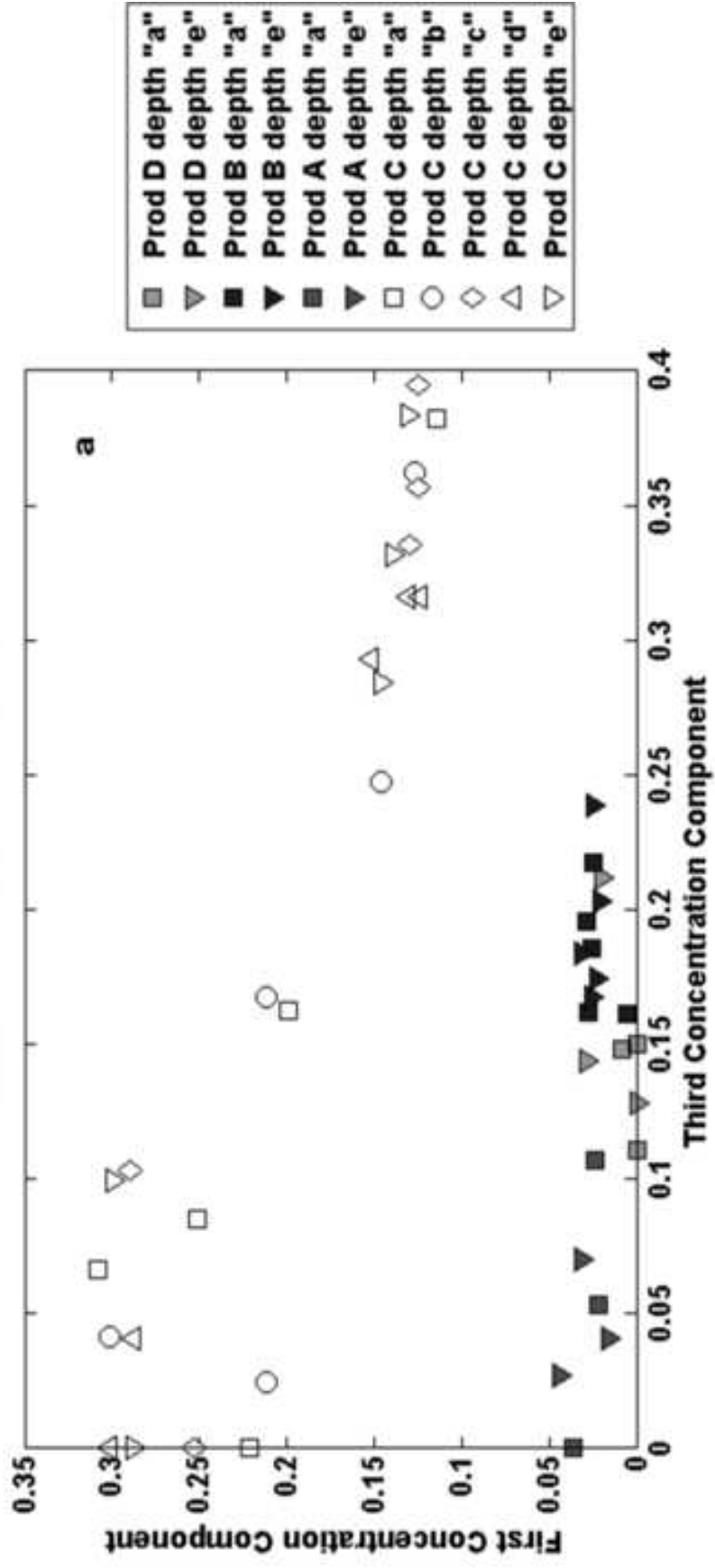


Figure 6b  
[Click here to download high resolution image](#)

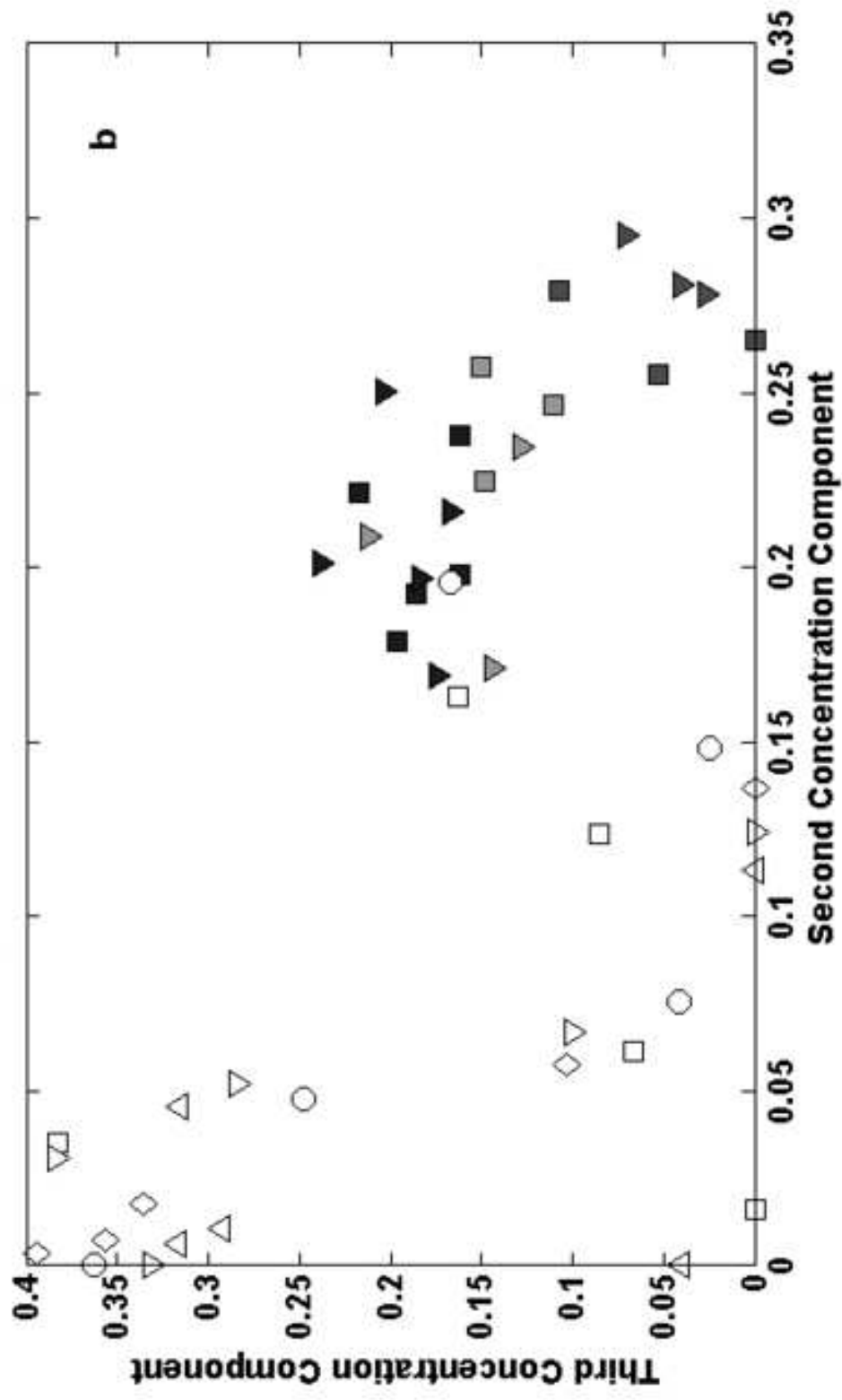


Figure 6c  
[Click here to download high resolution image](#)

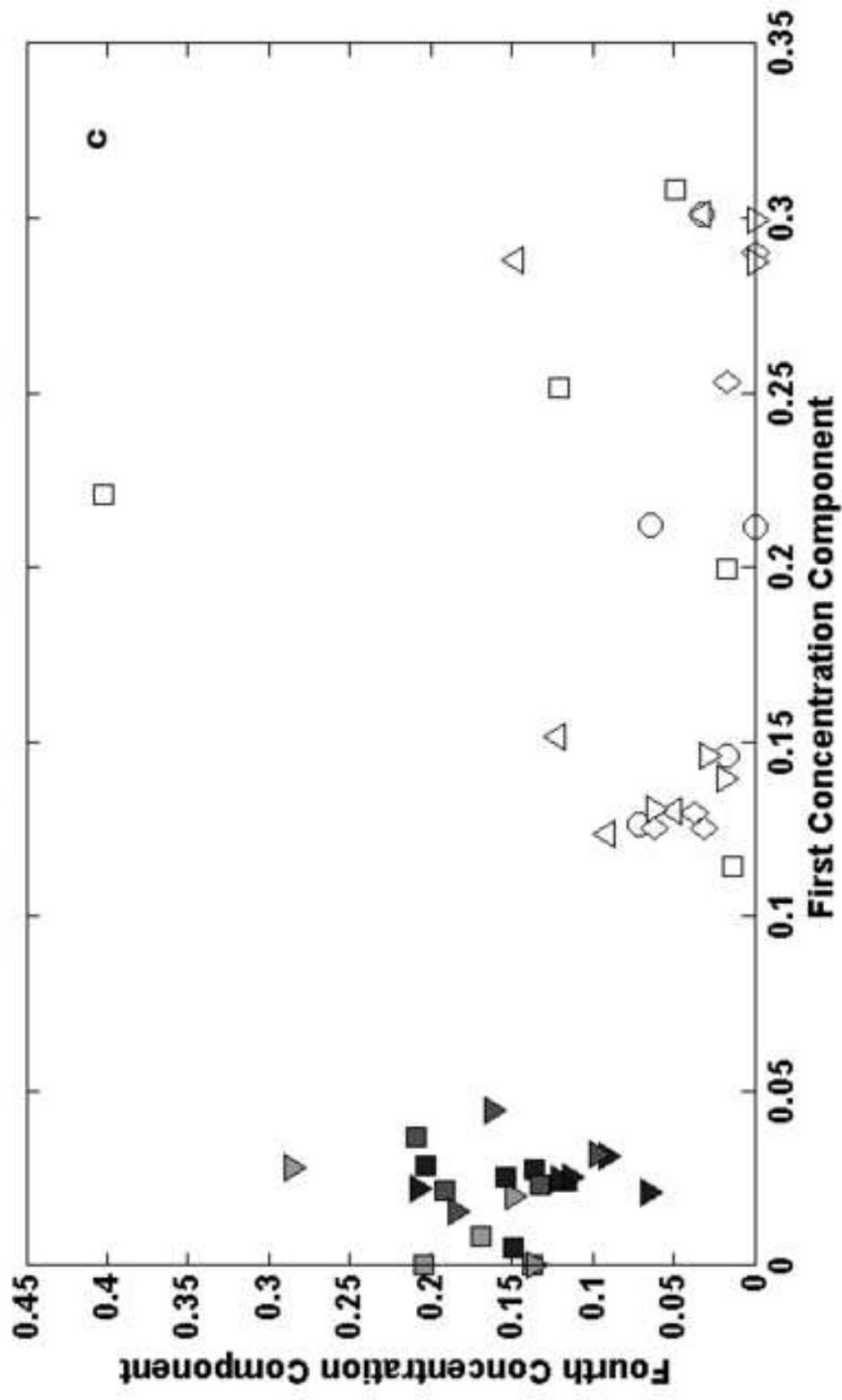
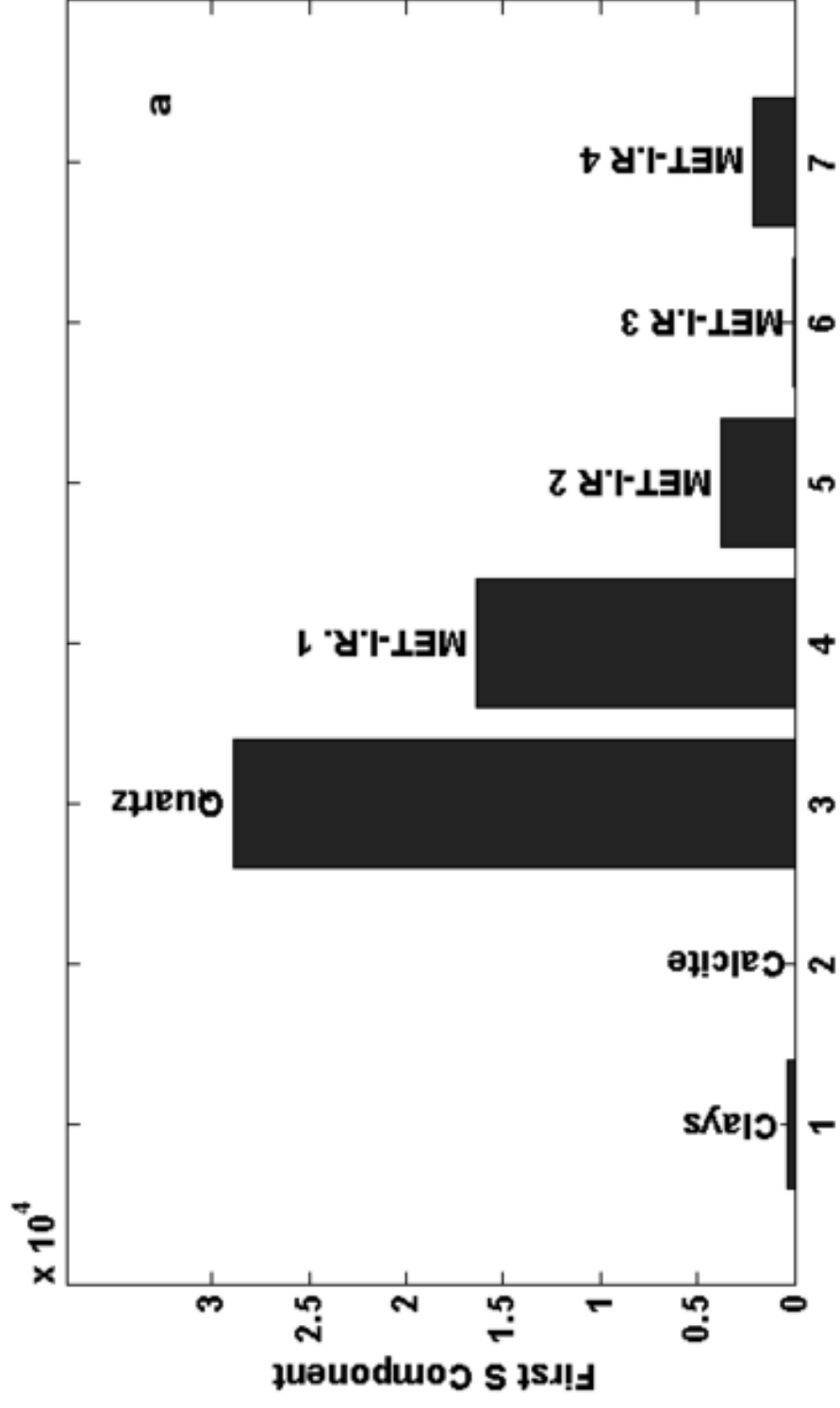


Figure 7a  
[Click here to download high resolution image](#)



a

Figure 7b  
[Click here to download high resolution image](#)

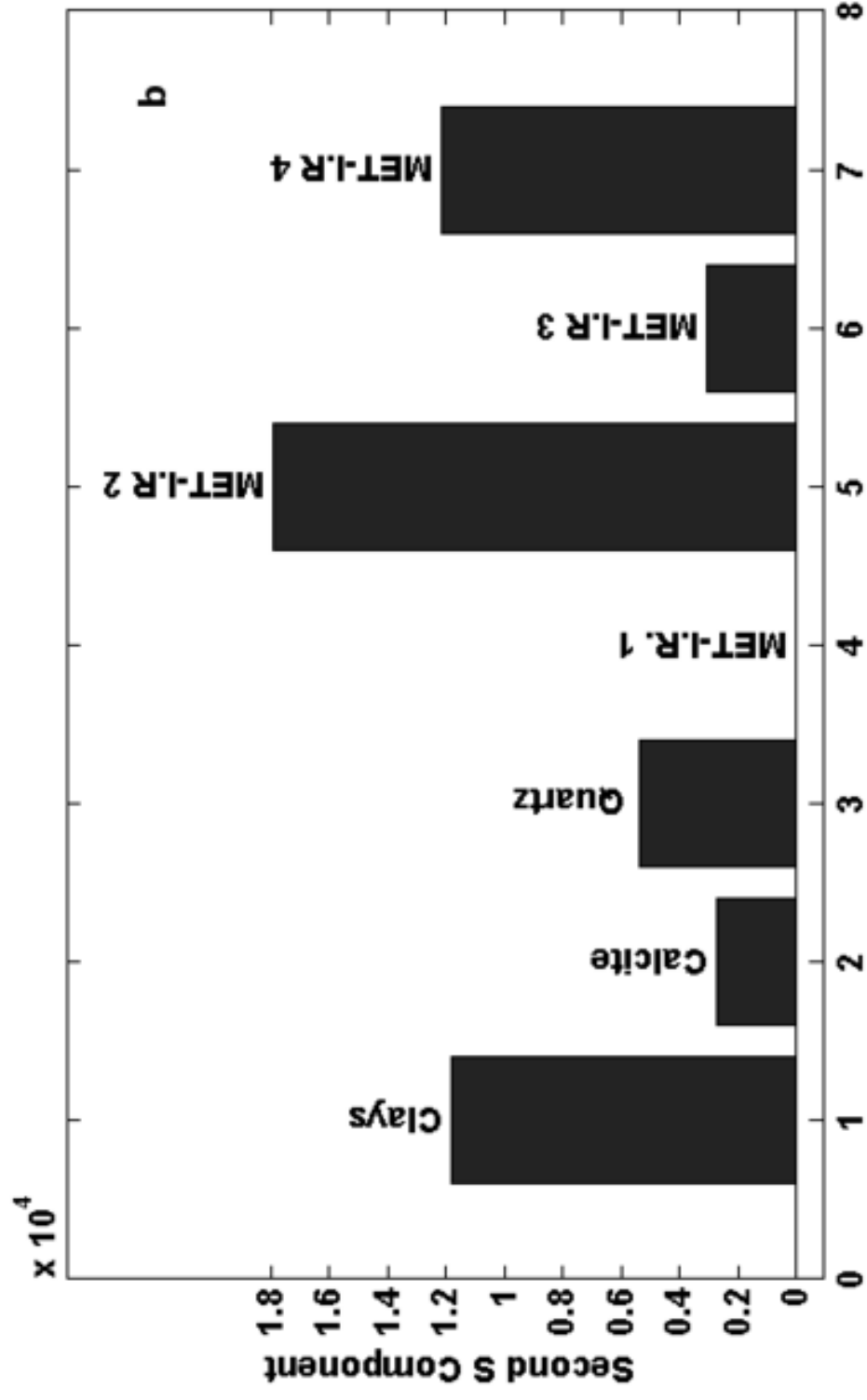


Figure 7c  
[Click here to download high resolution image](#)

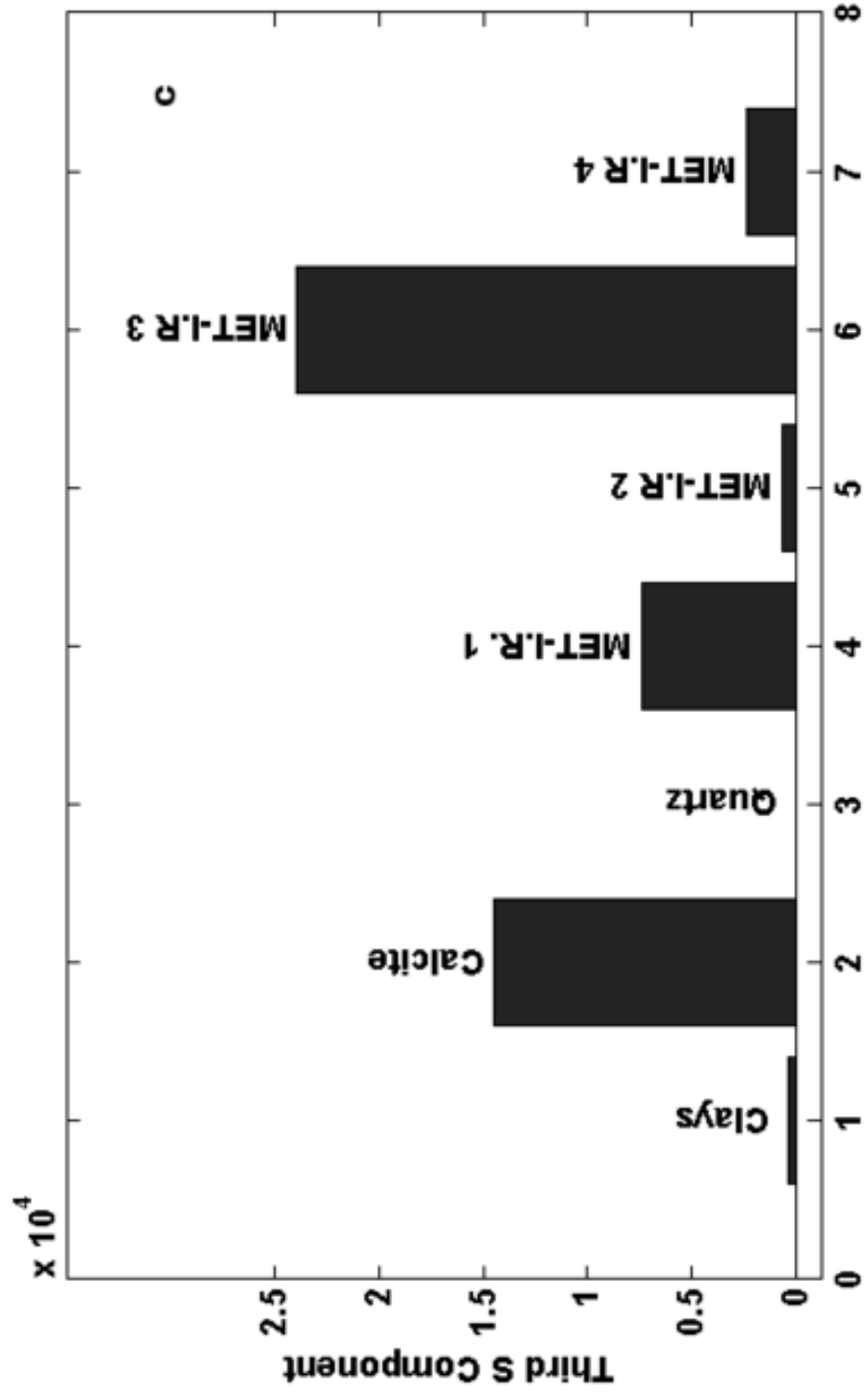




Figure 7d  
[Click here to download high resolution image](#)

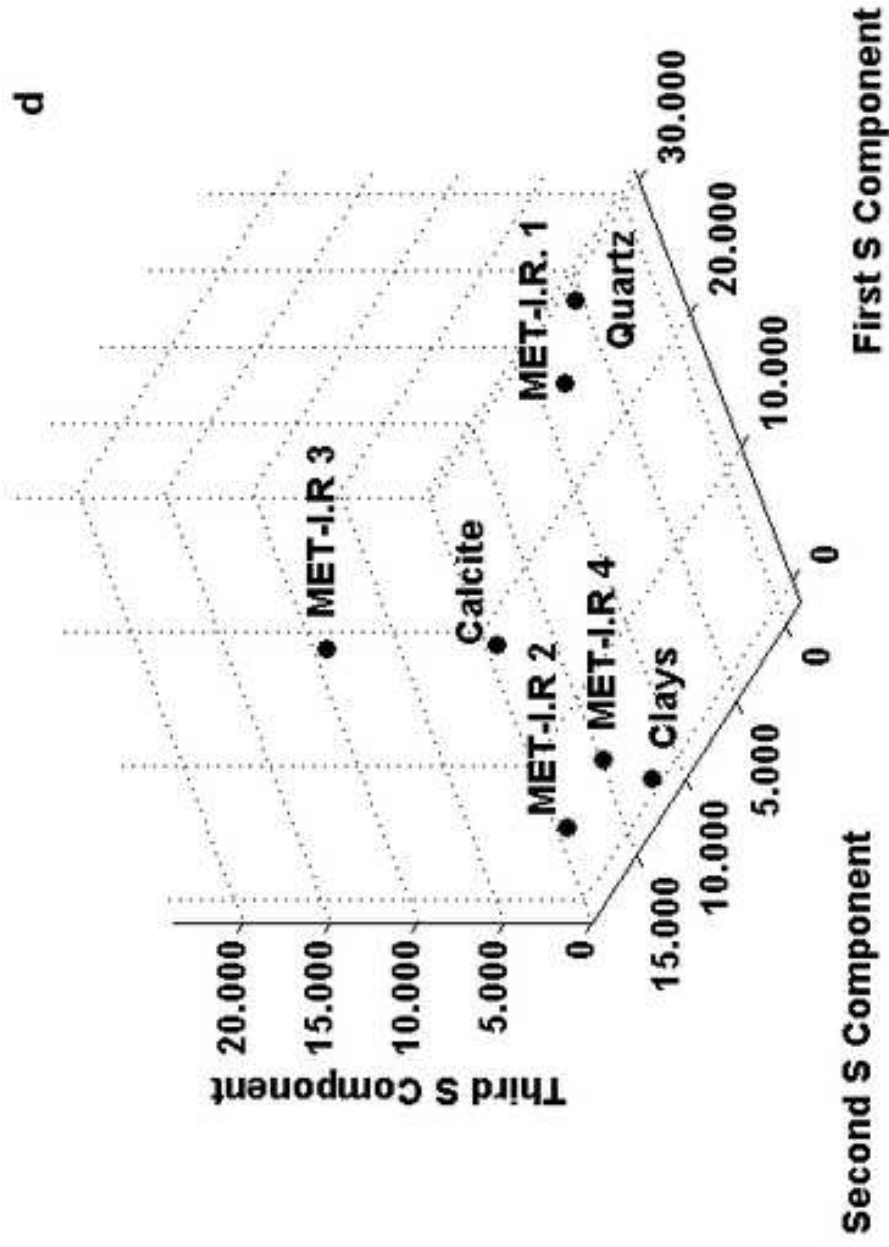


Figure 8a  
[Click here to download high resolution image](#)

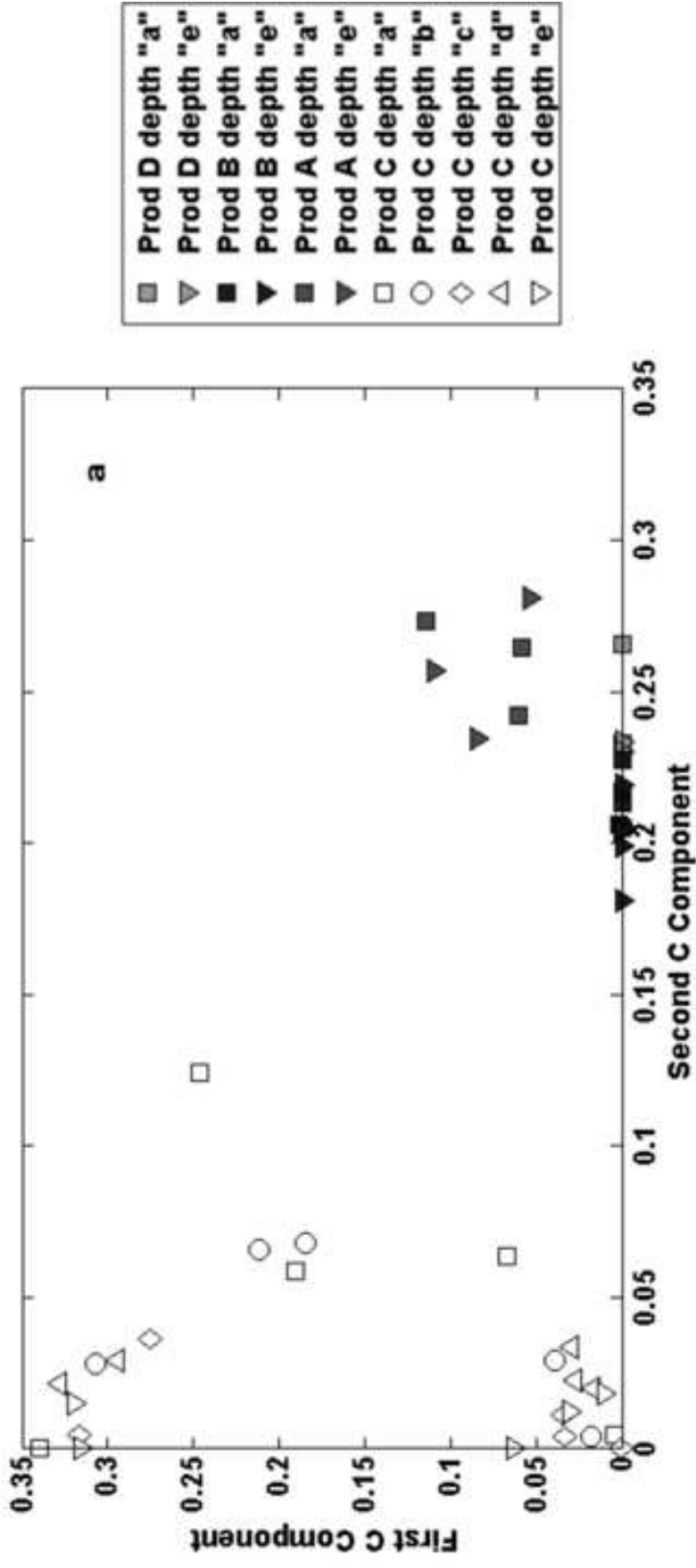


Figure 8b  
[Click here to download high resolution image](#)

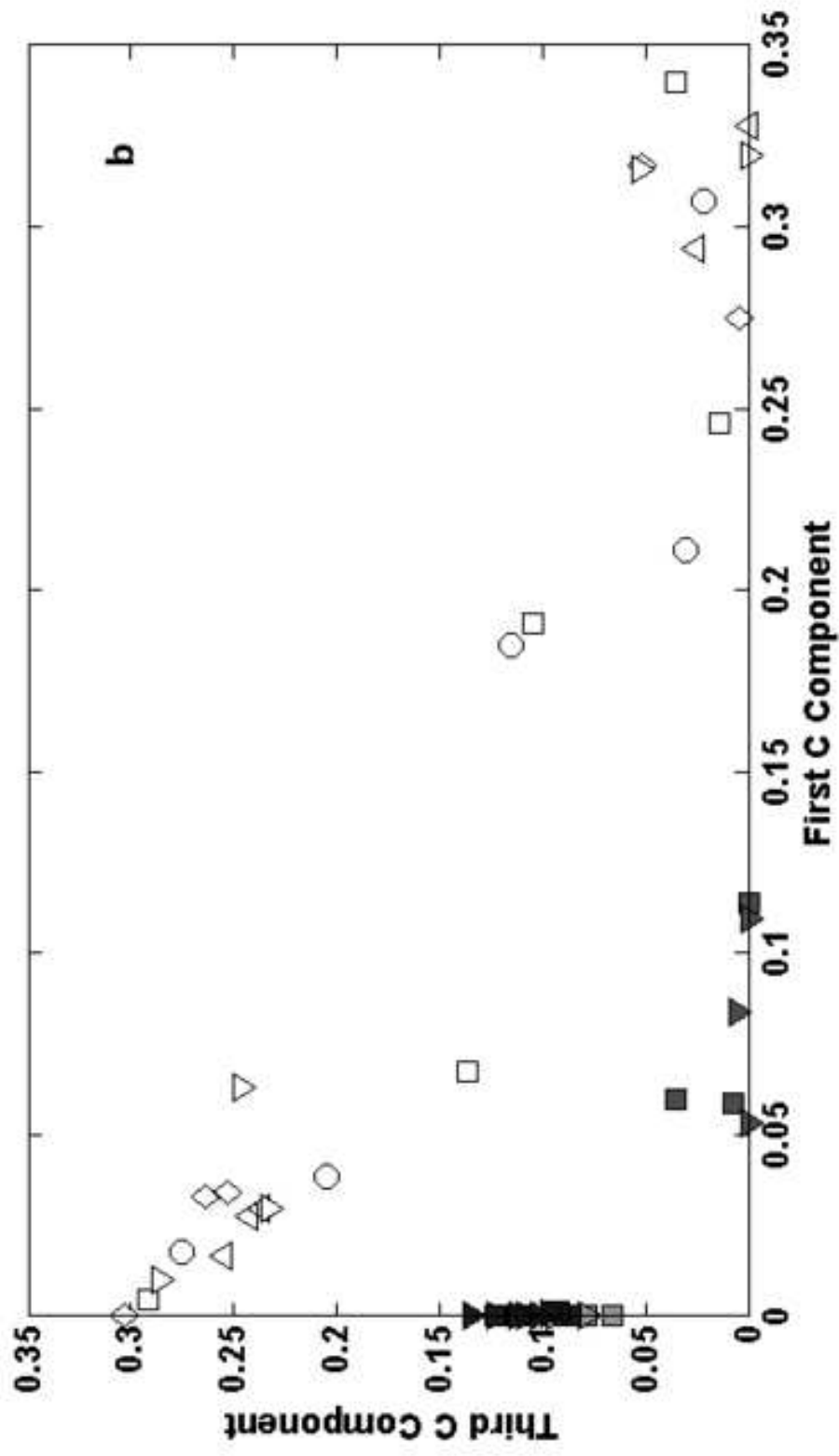


Figure 8c  
[Click here to download high resolution image](#)

